Impact of Smearing Techniques on Text line Localization of Kannada Historical Scripts

Vishwas H. S.
Dept of Electronics and Communication
Vidya Vikas Institute of Engineering
Mysore -570028

Bindu A.Thomas
Dept of Electronics and Communication
Vidya Vikas Institute of Engineering
Mysore-570028

ABSTRACT
Text line localization and segmentation is an important preprocessing stage in the context of document image analysis. Text lines must be localized first and then segmented in the logical order. The final recognition results are highly dependent on the results of text line segmentation. Historical documents have free form handwritten text and pose a great challenge for text line segmentation. The presence of connected and overlapping characters make the segmentation task more challenging. Smearing techniques have been conventionally used for the purpose of text line localization. In this paper, performance analysis of various smearing techniques is carried out on the text line localization of Kannada historical scripts.

General Terms
Text Line Segmentation, Historical Document Image Analysis

Keywords
Smearing Techniques, Run Length Smoothing Algorithm, Binary Transition Count Map, Adaptive Local Connectivity Map, touching / overlapping components

1. INTRODUCTION
Historical Documents are valuable source of documents for any country in terms of its cultural heritage. They contain valuable information about the past, pertaining to various fields either be medicine, astronomy or functioning of the society. Such documents need to be preserved electronically for the dissemination of the knowledge contained in them. Preserving them into an image format is not sufficient when recognition of handwritten text is needed for indexing and retrieval of such documents. Document image segmentation into its constituent parts such as Text lines, words and characters are important preprocessing stages and the final recognition results are highly dependent on these stages. A text line could be considered as a group of connected components that are adjacent or close to each other and corresponds to occurrence of text elements or it could be closed curve that represents the boundary of each text line [1].Several Techniques have been proposed in the literature for the segmentation of the documents into text lines. The projection profile based methods are well suited for machine printed documents, but are not suitable for handwritten documents with the presence of skew, touching and overlapping components. An improvement to projection profile analysis by dividing the document into small regions and obtaining the partial projection has also been described in the literature [1]. Hough transform based techniques provide good results when the text lines are along a same straight line, but the method fails when there is intra base line variability. Several Modifications to Hough transform has been proposed in [2]. Smearing based techniques are employed for both printed and handwritten documents. Grouping methods, also known as bottom up approaches are well known techniques in solving text line segmentation of printed documents. These methods group the Connected Components by considering the geometrical and topological characteristics. An approach using neighborhood connected component analysis is described in [3].

In [4], a piece-wise painting algorithm has been described by dividing the image into vertical strip. The map generated from the original image is subjected to mathematical and morphological operations, to obtain the line separators between the text lines. A similar approach has been proposed in [5] for the text line separation of Uyghur handwritten documents. Combination of morphological and smearing techniques has been used quite often in the literature for the segmentation of the unconstrained handwritten text lines. In [6] information based on foreground and background of the document image has been utilized for the separation of text lines. The technique employs run length smoothing algorithm and morphological operations. A morphological dilation based approach to segment handwritten Persian documents has been proposed in [7]. Cross counting technique has been employed in [8] [9] for getting a map of the original image. A graph cut algorithm or a suitable skeletonization algorithm has been used to segment out the touching parts in the mapped image. A layout and writer independent text line segmentation using a graph based approach is presented in [10]. The paper tries to solve the problem of touching and curvilinear text lines by localizing the text lines and segmenting them. Methods based on dividing the image into vertical strips and obtaining the map has been described in [11] [12] for segmenting complex Arabic historical documents.

In this paper performance analysis of Run Length Smoothing Algorithm, Binary transition Count Map and Adaptive Local Connectivity Map on the text line localization of the Kannada Historical Scripts is carried out. Section 2 details about Kannada historical scripts and challenges encountered in the text line segmentation process. Section 3 describes the three different smearing techniques employed in text line segmentation process and section 4 gives experimental results followed by conclusion.

2. KANNADA HISTORICAL SCRIPT
Kannada script is one of the earliest known scripts of India with a long historical heritage. Kannada script is evolved from the Brahmi Script in 3rd century B.C. The three evolutionary stages in Kannada script are Pre-old Kannada, Old Kannada and Modern Kannada. The documentation of Pre-old and old Kannada scripts was on the stones, palm leaves, copper plates. The writing styles of Historical scripts vary widely from the present day modern Kannada script [16]. A huge amount of Kannada historical documents are available in universities, museums and in archeological departments. These documents need to be preserved digitally for future generations and exploration of the data.

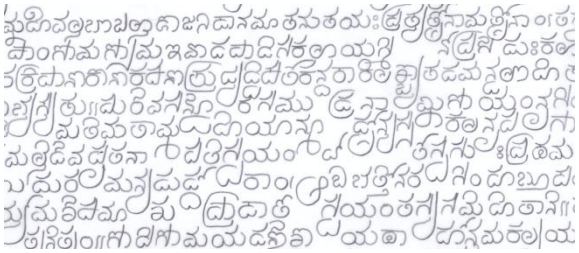

Figure1: A sample image of Kannada Historical Script belonging to Hoysala period

In this paper documents belonging to the Hoysala dynasty period are considered. Most of the documents pertaining to this dynasty period are found on the stones, walls of temple as inscriptions. The writing style is characterized by curved shape characters. The character belonging to Hoysala dynasty script resembles more of the modern day Kannada script. A sample document image is shown in Fig.1.

2.1 Segmentation Challenges

Historical document written in an unconstrained manner pose several challenges for the text line segmentation process. The main challenges arise from the existence of overlapping and/or touching lines, the variable character size and narrow line spacing between the text lines [10]. In specific, Kannada historical scripts present specific challenges when compared with other types of historical documents. The presence of long ascenders and descenders resulting from compound characters interfering between the characters in the neighboring text lines, overlapping of compound characters from the neighboring text lines make the process of text line localization and segmentation more challenging.

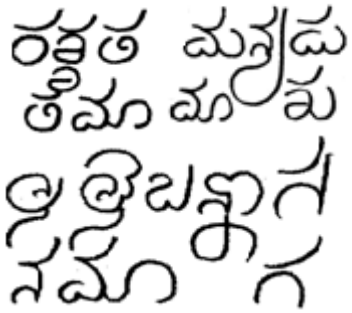

Figure 2: Challenges in the Text line Localization

3. SMEARING TECHNIQUES

Smearing techniques try to enhance the text area by creating homogeneous regions along the text locations. This section describes three main types of smearing algorithms namely Run Length Smearing Algorithm, Binary transition count map and Adaptive local connectivity map and their application on Kannada Historical Scripts.

3.1 Run Length Smoothing Algorithm

Run Length smoothing Algorithm is a popular technique for block segmentation in document images. They detect long vertical and horizontal white lines. This property can be utilized to localize text lines in document images. The basic RLSA [15] is applied on to a binary sequence in which white pixels are represented by 0's and black pixel by 1's.

Let the original binary image be denoted by I and output image by U. The RLSA algorithm transforms binary sequence in I to an output sequence with the following condition that 0's in I are changed to 1's in U if the number of horizontal adjacent 0's is less than or equal to predefined limit W_{th} and 1's in I are unchanged in U. Here W_{th} is the threshold calculated based on the average character width. The threshold is set to three times the average character width. Since the distance between neighboring characters along the same line will be less than the distance between the text lines, a horizontal RLSA can provide text line localization. In general, RLSA replaces a sequence of background pixels between two foreground pixels in a specified direction. When a binarization is available for a document, extraction of the text lines can be done by a connected component collection and grouping.

3.2 Binary Transition Count Map

In this technique, a map is generated from the original binary image by counting the transition from 0's to 1's and from 1's to 0's. A suitable sliding window is located on each pixel, counts the transition within the window limit and pixel location is placed with the count value. As a result, the output mapped image will be a gray scale image. This gray scale image is converted to binary image using a suitable threshold to obtain a suitable connectivity map. This technique can be used to measure the document complexity [13].

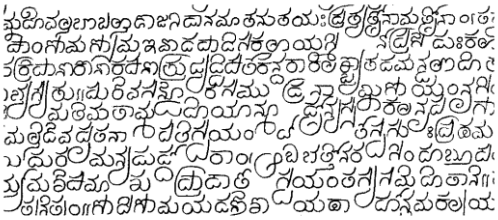

Figure 3: Binarized Image of Figure 1

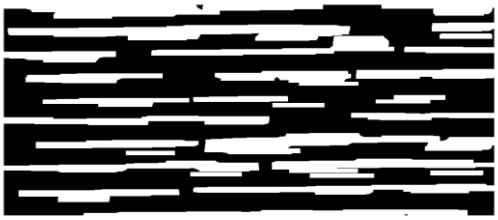

Figure 4: Map obtained after applying RLSA algorithm to Figure 3

Let I_b be the input binary image. Then the transition count map can be calculated using equation 1

$$U_b = \sum_{j=y-d}^{y+d} Ib(x,j) \cdot Ib(x,j+1) \qquad 1$$

Where U is the mapped output gray scale image and d is the size of the sliding window in horizontal direction. The width of the sliding window is set to average character width size. The sliding window is capable to capture the binary transitions within its window size and obtain the count. As a result the sliding window is able to capture the local features around the pixel. The resulting gray scale map is converted to a binary image using Otsu's thresholding algorithm. The binary gray map is shown in fig.4

Figure 5: Map obtained by applying transition count map algorithm to Figure 3

3.3 Adaptive Local Connectivity Map

Adaptive Local Connectivity Map [14] tries to capture the connectivity features around the pixels in the document images. A sliding window is used to scan the document image in a raster scan order. Within each window cumulative sum is calculated around the pixel and the sum is placed in the coordinates of the positioned pixel. A higher value in the map means that pixel is in the dense text region. Hence the mapped image is a gray scale image. Unlike transition count map, connectivity map operates on gray scale document images. The results are then thresholded to obtain a suitable binary image indicating the text line localized regions. The Sliding window is chosen with size twice the average character height.

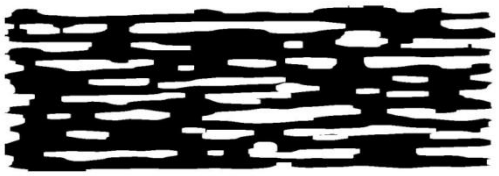

Figure 6: Map obtained by applying Connectivity Map algorithm to Figure 1

4. EXPERIMENTAL RESULTS

Smearing techniques explained in this paper was implemented using Matlab12.0Ra and tested on Kannada historical database consisting of around 25 images with approximately 400 text lines with around 16 lines per image. The database was created by taking photographs of documents written on stones from various historical places like Somnathpura, Belur, Halebidu and Shravanabelagola of Hassan district, Karnataka state. The document images contained only textual information and all documents are single columned having touching and overlapping components.

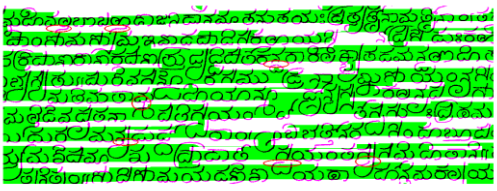

Figure 7: RLSA map overlapped on the original image also indicating failure cases in red color circles

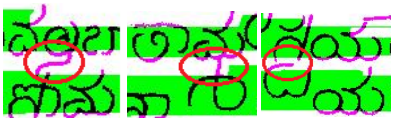

Figure 8: Red Colored circles indicate localization Failure due to RLSA

From the experimental results it is found that RLSA algorithm fails to localize the text lines in most of the cases, where there is overlapping and touching between the characters in the neighboring text lines. This is indicated in Fig.8 with the red circles marked. In the first case of Fig.8 an overlapping situation is indicated. In the second and third images of Fig.8 compound characters interfering from the neighboring text lines are shown. In all the three cases RLSA algorithm fails to localize the overlapping and interfering components from the neighboring text lines. Hence RLSA may not be a suitable technique when there is a need to localize text lines in the

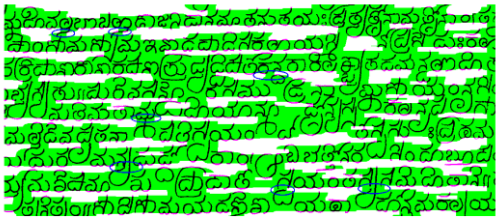

Figure 9: Overlapping of Transition Count Map on original image.

Figure 10: Blue colored circles indicate success over RLSA

presence of touching and overlapping components. But binary transition Count Map technique provided good result when compared with the RLSA technique. The failures encountered in the case of run length technique were avoided in the transition count map method and provided good localization results. From Figure.10 it can be clearly observed that the overlapping and interfering compound characters from the neighboring text lines have been clearly localized.

Figure 11: Overlap of Adaptive Local Connectivity Map on Original Image

From the ALCM results, it can be noticed that it provides best localization result when compared to the rest of the other methods. Even in some cases of transition count map there were results of false localization. The neighboring text lines were falsely connected in the cases of narrow text line spaced regions by creating a false localization. But these situations were avoided in connectivity map technique. As a comparison of Fig 9 and Fig 11 the connectivity map is less densely condensed when compared with the transition count technique and provides a more accurate text localization result. This comes from the fact that connectivity technique can capture the local features very well and can localize text regions in the presence

of touching and overlapping components. The results of localization depend highly on the size of the sliding window taken. By manually evaluating the results on the datasets, localization efficiencies obtained is shown in table below

Table 1: Text Line Localization Efficiency

Method	Localization efficiency
RLSA	60%
Binary Transition Count Map	83%
Adaptive Local Connectivity Map	90%

5. CONCLUSION AND FUTURE WORK

In this paper, performance analysis of three smearing techniques namely Run Length Smoothing Algorithm, Binary Transition Count Map and Adaptive Local Connectivity Map on the text line localization of Kannada Historical Scripts is carried out. From the experimental results, it is observed that run length smoothing algorithm failed to localize touching and overlapping components and yielded very low localization results. Transition Count Map technique yielded good results in comparison with run length smoothing technique in the presence of touching and overlapping components. Adaptive Local Connectivity Map was able to capture local features effectively and localize text lines effectively and yielded good results in comparison to other two methods. Suitable segmentation algorithm can be developed for text line localized images, to get better segmentation results.

6. REFERENCES

[1] Vassilis Katsouros and Vassilis Papavassiliou, "Segmentation of handwritten Document images into text lines", Institute for Language and Speech Processing / R.C."Athena" Greece.

[2] G.Louloudis, B.Gatos, C.Halatsis "Text Line Detection in Unconstrained Handwritten Documents Using a Block-Based Hough Transform Approach" *International Conference on Document Analysis and Recognition*, 2007.

[3] Abhishek Khandelwal, Pritha Choudry, Ram Sarkar, Subhadip Basu, Mita Nasipuri, Nibaran Das , " Text Line Segmentation for Unconstrained Handwritten Document Images using Neighborhood Connected Component Analysis " , *Third International Conference on Pattern Recognition and Machine Intelligence*, Volume 5909, 2009, pp 369-374

[4] Alireza Alaei, and Nagabhushan, P. and Umapada Pal, "Piece-wise painting technique for line segmentation of unconstrained handwritten text: a specific study with Persian text documents", *Pattern Analysis and Application*, 14 (4). pp. 381-394. 2011

[5] Kamil Moydin, Yi Xiaofong and Askar Hamdulla "Connected Component feature Analysis based Handwritten Uyghur Text Lines Detection and Separation Algorithm" *International Journal of Signal Processing, Image Processing and Pattern Recognition*, Vol.8, No.3 (2015), pp.291 -302

[6] Partha Pratim Roy, Umapada Pal, Joseph Llados, "Morphology based Handwritten Line Segmentation Using Foreground and Background Information", *International Conference on Frontiers in Handwriting Recognition*, Montreal, Canada, pp.241-246, 2008

[7] Abdollah AmirKahni-Shahraki, Amir Ebrahimi Ghahnavieh, Seyyed Abdollah Mirmahdavi," A Morphological Approach to Persian Handwritten Text Line Segmentation" *16th International Conference on Computer Modeling and Simulation*, 2014

[8] A.Sanchez, P.D.Suarez, C.A.B.Mello, A.L.I.Oliveria and V.M.O.Alves, "Text Line Segmentation in Images of Handwritten Historical Documents", *First Workshop on Image Processing Theory, Tools and Applications*, Sousse, 2008

[9] Douglas J.Kennard, William A. Barrett, "Separating Lines of Text in Free-Form Handwritten Historical Documents", *Proceedings of the Second International Conference on Document Image Analysis for Libraries* (DIAL), 2006

[10] David Fernandez-Mota, Joseph Llados, Alicia Fornes "A Graph based Approach for Segmenting Touching Lines in Historical Handwritten Documents", *International Journal of Document Analysis and Recognition (IJDAR),* 2014

[11] Wafa Boussella, Abderrazak Zahou, Haikal Elabed, Abdellatif Benabdelhafid, Adel Adimi," Unsupervised Block Covering Analysis for Text-Line Segmentation of Arabic Ancient Handwritten Document Images ", *International Conference on Pattern Recognition*, Istanbul 2010

[12] Abderrazak Zahour, Brunco Taconet, Laurence Likforman-Sulem, Wafa Boussellaa, " Overlapping and Multi - touching text line segmentation by Block Covering analysis " *Pattern Analysis and Applications*, Volume 12, Issue 4, pp335-351, 2009

[13] T.Akiyama and N.Hagita. Automated entry system for printed documents. *Pattern Recognition*, 23(11):1141-1151, 1990

[14] Z.Shi, S.Setlur and V.Govindaraju. Text extraction from gray Scale Historical document images using adaptive local connectivity map. *In Proceedings of 8th International Conference on Document Image Analysis and Recognition*, 2005

[15] K.Y.Wong, R.G Casey and F.M.Wahl,"Document analysis system," *IBM J.Res. Devel.* Vol.26, No.6, 111) 647-656, 1982.

[16] Dr.M.G Manjunath and G.K Devarajaswamy,"Kannada Lipi Vikasa", Yuvasadhane, Bengaluru.